

# Negative binomial factor regression with application to microbiome data analysis

Aditya K. Mishra<sup>1</sup>, Christian L. Müller

*Center for Computational Mathematics, Flatiron Institute, NY 10010*

<sup>1</sup>amishra@flatironinstitute.org

November 28, 2021

## 1 Supplementary Materials

### 1.1 Estimation Procedure for NB-RRR

We follow the framework of mixed-reduced-rank-regression (M-RRR) [Luo et al., 2018] to compute the parameters estimate of NB-RRR by solving the optimization problem (8). Our iterative procedure alternate between  $\mathbf{C}$ -step,  $\beta$ -step and  $\Phi$ -step to update the parameters  $\mathbf{C}$ ,  $\beta$  and  $\Phi$ , respectively, until convergence.

**C-step** For fixed  $\{\beta, \Phi\}$ , we denote  $\Theta$ , a function of  $\mathbf{C}$ , as  $\Theta(\mathbf{C})$ . Suppose differentiable  $\mathcal{L}(\Theta(\mathbf{C}), \Phi)$  is L-Lipschitz continuous gradient for some  $L_c$  i.e.,  $\|\nabla\mathcal{L}(\Theta(\check{\mathbf{C}})) - \nabla\mathcal{L}(\Theta(\mathbf{C}))\| \leq L_c\|\check{\mathbf{C}} - \mathbf{C}\|$  for any conformable  $\check{\mathbf{C}}$ . The statement holds for any  $\sup_{\mathbf{C}} \|\nabla^2\mathcal{L}(\Theta(\mathbf{C}))\| \leq L_c$ . Using the result, we majorize  $\mathcal{L}(\Theta(\mathbf{C}), \Phi)$  by a convex surrogate at a given  $\check{\mathbf{C}}$  as

$$\mathcal{L}(\Theta(\mathbf{C}), \Phi) \leq M(\mathbf{C}; \check{\mathbf{C}}) = \mathcal{L}(\Theta(\check{\mathbf{C}}), \Phi) + \text{tr}(\nabla\mathcal{L}(\Theta(\check{\mathbf{C}}), \Phi))^T\{\mathbf{C} - \check{\mathbf{C}}\} + L_c\|\check{\mathbf{C}} - \mathbf{C}\|^2/2. \quad (1)$$

Let us denote the current estimate of  $\mathbf{C}$  by  $\check{\mathbf{C}}$ . We update the parameter  $\mathbf{C}$  as

$$\bar{\mathbf{C}} \equiv \arg \min_{\mathbf{C}} M(\mathbf{C}; \check{\mathbf{C}}), \quad \text{s.t.} \quad \text{rank}(\mathbf{C}) \leq r.$$

The unique optimal solution is given by  $\bar{\mathbf{C}} = \mathbb{T}^{(r)}(\check{\mathbf{C}} - \nabla\mathcal{L}(\Theta(\check{\mathbf{C}}), \Phi)/L_c)$  where  $\mathbb{T}^{(r)}(\mathbf{M})$  extracts  $r$  singular value decomposition (SVD) components of matrix  $\mathbf{M}$ . In the parametric formulation (8), we relate  $\Theta(\mathbf{C})$  to the linear predictor  $\eta(\mathbf{C})$  as  $g(b'(\Theta(\mathbf{C}))) = \eta(\mathbf{C})$ . Interchangeably, we express the  $\mathcal{L}(\Theta(\mathbf{C}), \Phi)$  as function of  $\eta(\mathbf{C})$  as  $\mathcal{L}(\eta(\mathbf{C}), \Phi)$ . To compute  $L_c$ , we define

$$\sup_{\mathbf{C}} \|\nabla^2\mathcal{L}(\Theta(\mathbf{C}))\| = \max_{1 \leq j \leq q} \sup_{\mathbf{C}_{.j}} \|\nabla^2\mathcal{L}(\Theta(\mathbf{C}_{.j}))\|,$$

where  $\nabla^2\mathcal{L}(\Theta(\mathbf{C}_{.j})) = \mathbf{X}^T \text{diag}[\nabla^2_{.j}\mathcal{L}(\eta)]\mathbf{X}$  and  $\nabla^2\mathcal{L}(\eta) = [\nabla^2\mathcal{L}(\eta_{ij})]_{n \times q}$  such that

$$\nabla^2\mathcal{L}(\eta_{ij}) = (y_{ij} + 1) \frac{\phi_j \exp \eta_{ij}}{(\phi_j + \exp \eta_{ij})^2} \leq \frac{y_{ij} + 1}{4} \implies L_c = \max_{1 \leq j \leq q} \|\mathbf{X}^T \text{diag}(\mathbf{Y}_{.j} + 1)\mathbf{X}\|/2. \quad (2)$$

**$\beta$ -step:** For fixed  $\mathbf{C}$  and  $\Phi$ , we denote  $\Theta$ , a function of  $\beta$ , as  $\Theta(\beta)$ . In terms of  $\beta$ , suppose  $\mathcal{L}(\Theta(\beta), \Phi)$  is L-Lipschitz continuous gradient for some  $L_b$ . Following the  $\mathbf{C}$ -step, we majorize the objective function  $\mathcal{L}(\Theta(\beta), \Phi)$  at the current parameter  $\check{\beta}$  as

$$\mathcal{L}(\Theta(\beta), \Phi) \leq K(\beta; \check{\beta}) = \mathcal{L}(\Theta(\check{\beta}), \Phi) + \nabla\mathcal{L}(\Theta(\check{\beta}), \Phi)^T\{\beta - \check{\beta}\} + \frac{L_b}{2}\|\check{\beta} - \beta\|^2, \quad (3)$$

and update the parameter with the solution of the optimization problem

$$\bar{\beta} \equiv \arg \min_{\beta} K(\beta; \check{\beta}).$$

The unique optimal solution is given by  $\bar{\beta} = \check{\beta} - \nabla \mathcal{L}(\Theta(\check{\beta}), \Phi) / s_b$ . Following **C**-step, we compute  $L_b = \max_{1 \leq j \leq q} \|\mathbf{Z}^T(\mathbf{Y}_{\cdot j} + 1)\mathbf{Z}\|/2$ .

**$\Phi$ -step:** For fixed **C** and  $\beta$ , we apply Newton-Raphson [R Core Team, 2019] to update the current estimate of  $\Phi$  such that the negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  decreases; see Section 1.5 for more details on the update rule.

## 1.2 Proof of Theorem 1

We prove the monotone decreasing likelihood property of Algorithm 1 estimating parameters of the NB-RRR. Let us denote the parameters set in the  $t$ th iteration by  $\mathbf{L}^{(t)} = (\mathbf{C}^{(t)}, \beta^{(t)}, \Phi^{(t)})$ .

**C-step** With  $\{\beta^{(t)}, \Phi^{(t)}\}$  fixed, we denote the likelihood as function of **C** by  $\mathcal{L}(\Theta(\mathbf{C}), \Phi^{(t)})$ . Using the result in equation (1), for  $t$ th iteration, we have

$$\mathcal{L}(\Theta(\mathbf{C}), \Phi^{(t)}) \leq M(\mathbf{C}; \mathbf{C}^{(t)}) \quad \forall \quad \mathbf{C} \in \mathbb{R}^{p \times q}.$$

Let us denote the optimal solution minimizing the surrogate function  $M(\mathbf{C}; \mathbf{C}^{(t)})$  in the **C**-step by  $\mathbf{C}^{(t+1)}$ . For  $\mathbf{C}^{(t+1)}$ , we write the above result as

$$\mathcal{L}(\Theta(\mathbf{C}^{(t+1)}), \Phi^{(t)}) \leq M(\mathbf{C}^{(t+1)}; \mathbf{C}^{(t)}) \leq M(\mathbf{C}^{(t)}; \mathbf{C}^{(t)}) = \mathcal{L}(\Theta(\mathbf{C}^{(t)}), \Phi^{(t)}).$$

**$\beta$ -step** With  $\{\mathbf{C}^{(t+1)}, \Phi^{(t)}\}$  fixed, we denote the likelihood as function of  $\beta$  given by  $\mathcal{L}(\Theta(\beta), \Phi^{(t)})$ . Using the result in equation (3), for  $t$ th iteration, we have

$$\mathcal{L}(\Theta(\beta), \Phi^{(t)}) \leq K(\beta; \beta^{(t)}) \quad \forall \quad \beta \in \mathbb{R}^{c \times q}.$$

Let us denote the optimal solution minimizing the surrogate function  $K(\beta; \beta^{(t)})$  in the  $\beta$ -step by  $\beta^{(t+1)}$ . For  $\beta^{(t+1)}$ , we write the above result as

$$\mathcal{L}(\Theta(\beta^{(t+1)}), \Phi^{(t)}) \leq K(\beta^{(t+1)}; \beta^{(t)}) \leq K(\beta^{(t)}; \beta^{(t)}) = \mathcal{L}(\Theta(\beta^{(t)}), \Phi^{(t)}).$$

Finally, in  $\Phi$ -step, with fixed  $\mathbf{C}^{(t+1)}$  and  $\beta^{(t+1)}$ , we minimize the negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  with respect to  $\Phi$  using Newton-Raphson [R Core Team, 2019]. The approach guarantees to have non-increasing likelihood function  $\mathcal{L}(\Theta, \Phi)$ .

Combining the result of the non-increasing negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  in **C**-step,  $\beta$ -step and  $\Phi$ -step, we prove the monotone decreasing likelihood property of Algorithm 1.

## 1.3 Estimation Procedure for Solving NB-CURE Problem

We follow the framework of generalize co-sparse factor regression (GO-FAR) [Mishra et al., 2020] to compute the parameters estimate of NB-CURE by solving the optimization problem (11). Our iterative procedure alternate between **u**-step, **v**-step,  $\beta$ -step and  $\Phi$ -step to update the parameters in blocks of  $(\mathbf{u}, d)$ ,  $(\mathbf{v}, d)$ ,  $\beta$  and  $\Phi$ , respectively, until convergence.

**u-step:** For fixed  $\{\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$  with  $\mathbf{v}^T \mathbf{v} = 1$ , we express the  $\boldsymbol{\Theta}$  as a function of the product variable  $\check{\mathbf{u}} = d\mathbf{u}$ , denoted by  $\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^T)$ , and the objective function in (11) as  $F_\lambda(\check{\mathbf{u}}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi})$ . In terms of  $\check{\mathbf{u}}$ , let us assume that the  $\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^T), \boldsymbol{\Phi})$  is L-Lipschitz continuous gradient for some  $L_u$ . Following the **C**-step in Section 1.1 for NB-RRR, we majorize the objective function (11) as

$$F_\lambda(\mathbf{a}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}) = \mathcal{L}(\boldsymbol{\Theta}(\mathbf{a}\mathbf{v}^T), \boldsymbol{\Phi}) + \rho_\lambda(\mathbf{a}\mathbf{v}^T; \mathbf{W}) \leq G_\lambda(\mathbf{a}; \check{\mathbf{u}}) \quad \forall \quad \mathbf{a} \in \mathbb{R}^p, \quad (4)$$

$$\text{where } G_\lambda(\mathbf{a}; \check{\mathbf{u}}) = \mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^T), \boldsymbol{\Phi}) + \nabla \mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^T), \boldsymbol{\Phi})^T (\mathbf{a} - \check{\mathbf{u}}) + \frac{L_u}{2} \|\check{\mathbf{u}} - \mathbf{a}\|^2 + \rho_\lambda(\mathbf{a}\mathbf{v}^T; \mathbf{W}),$$

and solve the optimization problem

$$\hat{\mathbf{u}} \equiv \arg \min_{\mathbf{a}} G_\lambda(\mathbf{a}; \check{\mathbf{u}}),$$

to update  $\check{\mathbf{u}}$ . Following Zou and Hastie [2005], the unique optimal solution is given by

$$\hat{\mathbf{u}} = \frac{\mathbf{S}(\check{\mathbf{u}} - \nabla \mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^T), \boldsymbol{\Phi})/L_u; \alpha \lambda \mathbf{v}^T \mathbf{w}^{(v)} \mathbf{w}^{(d)} \mathbf{w}^{(u)}/L_u)}{1 + 2\lambda(1 - \alpha) \|\mathbf{v}\|_2^2 / L_u}, \quad (5)$$

where  $\mathbf{S}(\mathbf{t}; \tilde{\lambda}) = \text{sign}(\mathbf{t})(|\mathbf{t}| - \tilde{\lambda})_+$  is the elementwise soft-thresholding operator on any  $\mathbf{t} \in \mathbb{R}^p$ . We utilize the equality constraint in equation (11) to recover the estimate of  $(d, \mathbf{u})$  from  $\hat{\mathbf{u}}$ . To compute  $L_u$ , we derive  $\frac{\partial \mathcal{L}}{\partial \check{\mathbf{u}}} = \mathbf{X}^T \nabla \mathcal{L}(\boldsymbol{\eta}) \mathbf{v} = \sum_{i=1}^n \mathbf{x}_i \sum_{k=1}^q \nabla \mathcal{L}(\eta_{ik}) v_k$  and  $\frac{\partial^2 \mathcal{L}}{\partial \check{\mathbf{u}} \partial \check{\mathbf{u}}^T} = \sum_{i=1}^n \mathbf{x}_i \left( \sum_{k=1}^q \nabla^2 \mathcal{L}(\eta_{ik}) v_k^2 \right) \mathbf{x}_i^T$  and utilize the result in (2) to upper bound the  $\sup_{\check{\mathbf{u}}} \left\| \frac{\partial^2 \mathcal{L}}{\partial \check{\mathbf{u}} \partial \check{\mathbf{u}}^T} \right\|$  as

$$L_u = \left\| \sum_{i=1}^n \mathbf{x}_i \left( \sum_{k=1}^q (y_{ik} + 1) v_k^2 \right) \mathbf{x}_i^T \right\| / 2 = \left\| \mathbf{X}^T \mathbf{X} + \sum_{i=1}^n \mathbf{x}_i \left( \sum_{k=1}^q y_{ik} v_k^2 \right) \mathbf{x}_i^T \right\| / 2.$$

**v-step** For fixed  $\{\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ , we express  $\boldsymbol{\Theta}$  as a function of the product variable  $\check{\mathbf{v}} = d\mathbf{v}$ , denoted by  $\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^T)$ , and the objective function in (11) as  $F_\lambda(\mathbf{u}, \check{\mathbf{v}}, \boldsymbol{\beta}, \boldsymbol{\Phi})$ . In terms of  $\check{\mathbf{v}}$ , suppose  $\mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^T), \boldsymbol{\Phi})$  is L-Lipschitz continuous gradient for some  $L_v$ . Again we majorize the objective function (11) as

$$F_\lambda(\mathbf{u}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\Phi}) = \mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\mathbf{b}^T), \boldsymbol{\Phi}) + \rho_\lambda(\mathbf{u}\mathbf{b}^T; \mathbf{W}) \leq H_\lambda(\mathbf{b}; \check{\mathbf{v}}) \quad \forall \quad \mathbf{b} \in \mathbb{R}^q, \quad (6)$$

$$\text{where } H_\lambda(\mathbf{b}; \check{\mathbf{v}}) = \mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^T), \boldsymbol{\Phi}) + \nabla \mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^T), \boldsymbol{\Phi})^T (\mathbf{b} - \check{\mathbf{v}}) + \frac{L_v}{2} \|\check{\mathbf{v}} - \mathbf{b}\|^2 + \rho_\lambda(\mathbf{u}\mathbf{b}^T; \mathbf{W}),$$

and solve the optimization problem

$$\hat{\mathbf{v}} \equiv \arg \min_{\mathbf{b}} H_\lambda(\mathbf{b}; \check{\mathbf{v}}),$$

to update  $\check{\mathbf{v}}$ . Following the **u**-step, the unique optimal solution is given by

$$\hat{\mathbf{v}} = \frac{\mathbf{S}(\check{\mathbf{v}} - \nabla \mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^T), \boldsymbol{\Phi})/L_v; \alpha \lambda \mathbf{u}^T \mathbf{w}^{(u)} \mathbf{w}^{(d)} \mathbf{w}^{(v)}/L_v)}{1 + 2\lambda(1 - \alpha) \|\mathbf{u}\|_2^2 / L_v}. \quad (7)$$

Again, we utilize the equality constraint (11) to recover the estimate of  $(d, \mathbf{v})$  from  $\hat{\mathbf{v}}$ . To compute  $L_v$ , we derive  $\frac{\partial \mathcal{L}}{\partial \check{\mathbf{v}}} = \left( \frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}} \right)^T \mathbf{X} \mathbf{u} = \sum_{i=1}^n \mathbf{x}_i^T \mathbf{u} \nabla \mathcal{L}(\eta_i)$  and  $\frac{\partial^2 \mathcal{L}}{\partial \check{\mathbf{v}} \partial \check{\mathbf{v}}^T} = \text{diag}[\mathbf{u}^T \mathbf{X}^T \text{diag}(\nabla^2 \mathcal{L}(\eta_{1.})) \mathbf{X} \mathbf{u}, \dots, \mathbf{u}^T \mathbf{X}^T \text{diag}(\nabla^2 \mathcal{L}(\eta_{.q})) \mathbf{X} \mathbf{u}]$ . We utilize the result in (2) and upper bound  $\sup_{\check{\mathbf{v}}} \left\| \frac{\partial^2 \mathcal{L}}{\partial \check{\mathbf{v}} \partial \check{\mathbf{v}}^T} \right\|$  as

$$L_v = \max_{1 \leq j \leq q} \mathbf{u}^T \mathbf{X}^T \text{diag}(\nabla^2 \mathcal{L}(\eta_{.j})) \mathbf{X} \mathbf{u} / 2 = \max_{1 \leq j \leq q} \mathbf{u}^T \mathbf{X}^T \text{diag}[\mathbf{Y}_{.j} + 1] \mathbf{X} \mathbf{u} / 2.$$

**$\beta$ -step:** For fixed  $\mathbf{C} = d\mathbf{u}\mathbf{v}^\top$  and  $\Phi$ , we follow  $\beta$ -step update presented in Section 1.1 of NB-RRR to update the parameter  $\beta$ .

**$\Phi$ -step:** For fixed  $\mathbf{C}$  and  $\beta$ , we apply Newton-Raphson [R Core Team, 2019] to update the current estimate of  $\Phi$  such that the negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  decreases; see Section 1.5 for more details on the update rule.

## 1.4 Proof of Theorem 2

We prove the monotone decreasing likelihood property of Algorithm 3 solving the NB-CURE optimization problem. Let us denote the parameters set in the  $t$ th iteration by  $\mathbb{L}^{(t)} = (d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \beta^{(t)}, \Phi^{(t)})$ .

**u-step** With  $\{\mathbf{v}^{(t)}, \beta^{(t)}, \Phi^{(t)}\}$  fixed, we update the block variables  $\{d^{(t)}, \mathbf{u}^{(t)}\}$  as  $\{\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)}\}$  using equation (5). Using the result in (4), we have

$$F_\lambda(\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t)}, \beta^{(t)}, \Phi^{(t)}) \leq G_\lambda(\check{\mathbf{u}}^{(t+1)}; \check{\mathbf{u}}^{(t)}) \leq G_\lambda(\check{\mathbf{u}}^{(t)}; \check{\mathbf{u}}^{(t)}) = F_\lambda(d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \beta^{(t)}, \Phi^{(t)}),$$

where  $\check{\mathbf{u}}^{(t)} = d^{(t)}\mathbf{u}^{(t)}$  and  $\check{\mathbf{u}}^{(t+1)} = \tilde{d}^{(t+1)}\mathbf{u}^{(t+1)}$ .

**v-step** With  $\{\mathbf{u}^{(t+1)}, \beta^{(t)}, \Phi^{(t)}\}$  fixed, we update the block variables  $\{d^{(t+1)}, \mathbf{v}^{(t)}\}$  as  $\{d^{(t+1)}, \mathbf{u}^{(t+1)}\}$  using equation (7). Using the result in (6), we have

$$F_\lambda(d^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)}, \beta^{(t)}, \Phi^{(t)}) \leq H_\lambda(\check{\mathbf{v}}^{(t+1)}; \check{\mathbf{v}}^{(t)}) \leq H_\lambda(\check{\mathbf{v}}^{(t)}; \check{\mathbf{v}}^{(t)}) = F_\lambda(\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t)}, \beta^{(t)}, \Phi^{(t)}),$$

where  $\check{\mathbf{v}}^{(t)} = \tilde{d}^{(t+1)}\mathbf{v}^{(t)}$  and  $\check{\mathbf{v}}^{(t+1)} = d^{(t+1)}\mathbf{v}^{(t+1)}$ .

**$\beta$ -step** With  $\{\mathbf{u}^{(t+1)}, d^{(t+1)}, \mathbf{v}^{(t+1)}, \Phi^{(t)}\}$  fixed, we update  $\beta^{(t)}$  as  $\beta^{(t+1)}$  by following  $\beta$ -step update presented in Section 1.1 of NB-RRR. On extending the non-decreasing result of  $\beta$ -step update presented in Section 1.2, we have

$$F_\lambda(d^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)}, \beta^{(t+1)}, \Phi^{(t)}) \leq F_\lambda(d^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)}, \beta^{(t)}, \Phi^{(t)}).$$

Finally, in  $\Phi$ -step, with fixed  $\mathbf{C}^{(t+1)} = d^{(t+1)}\mathbf{u}^{(t+1)}\mathbf{v}^{(t+1)\top}$  and  $\beta^{(t+1)}$ , we minimize the negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  with respect to  $\Phi$  using Newton-Raphson [R Core Team, 2019]. The approach guarantees to have non-increasing likelihood function  $\mathcal{L}(\Theta, \Phi)$ .

Combining the result of the non-increasing negative log-likelihood function  $\mathcal{L}(\Theta, \Phi)$  in **u-step**, **v-step**,  **$\beta$ -step** and  **$\Phi$ -step**, we prove the monotone decreasing likelihood property of Algorithm 3.

## 1.5 Newton-Raphson to update $\Phi$ :

For fixed  $\mathbf{C}$  and  $\beta$ , we have the mean parameter matrix as  $\boldsymbol{\mu} = g^{-1}(\mathbf{O} + \mathbf{X}\mathbf{C} + \mathbf{Z}\beta)$ . Following [Ismail and Jemain, 2007], we reparametrize  $\mathcal{L}(\Theta, \Phi)$  in terms of the inverse of the shape parameter by replacing  $\phi_j = 1/\alpha_j$  for  $j = 1, \dots, q$  and write the likelihood function as

$$\mathcal{L}(\Theta, \alpha) = \sum_{i=1}^n \left[ \left( \sum_{k=0}^{y_{ij}-1} \log \frac{k\alpha_j + 1}{\alpha_j} \right) + y_{ij} \log \alpha_j - \left( y_{ij} + \frac{1}{\alpha_j} \right) \log(1 + \mu_{ij}\alpha_j) \right],$$

Table S1: Simulation: model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in case of  $p = 100$  with negative binomial responses.

	M%	Er(C)	Er( $\Theta$ )	FPR	FNR	R%	r	Er( $\Phi$ )	Time(s)
NB-FAR	0	3.89 (0.97)	16.56 (2.07)	9.74 (4.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	0.13 (0.08)	157.66 (22.71)
NB-RRR	0	19.67 (2.18)	31.07 (1.39)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	0.36 (0.10)	48.49 (6.61)
GO-FAR	0	27.00 (7.01)	43.62 (5.60)	6.88 (5.32)	51.47 (29.14)	0.61 (1.83)	3.30 (0.68)	1.37 (0.00)	31.95 (4.74)
NB-GLM	0	237.57 (12.47)	93.87 (2.34)	100.00 (0.00)	0.00 (0.00)	51.72 (2.10)	29.29 (0.72)	4.54 (0.41)	74.62 (21.17)
NB-FAR	20	5.19 (1.73)	18.53 (2.49)	10.84 (4.66)	0.59 (1.03)	0.00 (0.00)	3.00 (0.00)	0.15 (0.08)	162.06 (21.45)
NB-RRR	20	22.22 (2.99)	33.04 (1.48)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	0.43 (0.10)	49.63 (7.10)
GO-FAR	20	29.66 (8.04)	46.98 (7.82)	6.86 (5.67)	50.26 (30.17)	0.00 (0.00)	3.00 (0.00)	1.37 (0.00)	35.01 (5.64)
NB-GLM	20	294.11 (16.61)	103.86 (2.31)	99.87 (0.32)	0.00 (0.00)	55.90 (2.05)	29.06 (1.09)	6.04 (0.62)	125.09 (59.07)

where  $\boldsymbol{\alpha} = \text{diag}(\alpha_1, \dots, \alpha_q)$ . Based on the Newton-Raphson method, we update the current estimate  $\alpha_j$  as  $\hat{\alpha}_j = \alpha_j - \left(\frac{\partial^2 \mathcal{L}}{\partial \alpha_j^2}\right)^{-1} \frac{\partial \mathcal{L}}{\partial \alpha_j}$ . We write the first derivative and second derivative of  $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha})$  with respect to any  $\alpha_j$  as

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha})}{\partial \alpha_j} &= \sum_{i=1}^n \left[ \left( \sum_{k=0}^{y_{ij}-1} \frac{k}{k\alpha_j + 1} \right) - \frac{\mu_{ij} y_{ij} \alpha_j + 1}{\alpha_j \mu_{ij} \alpha_j + 1} + \frac{1}{\alpha_j^2} \log(1 + \mu_{ij} \alpha_j) \right] \\ \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha})}{\partial \alpha_j^2} &= \sum_{i=1}^n \left[ \left( \sum_{k=0}^{y_{ij}-1} \frac{-k^2}{(k\alpha_j + 1)^2} \right) - \frac{2}{\alpha_j^3} \log(1 + \mu_{ij} \alpha_j) + \frac{\mu_{ij} [2 + \mu_{ij} \alpha_j (3 + y_{ij} \alpha_j)]}{\alpha_j^2 (\mu_{ij} \alpha_j + 1)^2} \right]. \end{aligned}$$

## 1.6 Some partial derivative of $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ for fixed $\boldsymbol{\Phi}$

For the ease of notation, we write  $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$  as  $\mathcal{L}(\boldsymbol{\Theta})$ . Here  $\boldsymbol{\Theta}$  is linked to linear predictor  $\boldsymbol{\eta}$  as  $g(b'(\boldsymbol{\Theta})) = \boldsymbol{\eta}$  and

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}) &= -\text{tr}(\mathbf{Y}^T \boldsymbol{\Theta}) + \text{tr}(\mathbf{J}^T \mathbf{B}(\boldsymbol{\Theta})) + \text{const}, \\ \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \theta_{ij}} &= -y_{ij} + b'(\theta_{ij}) \implies \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \eta_{ij}} = (-y_{ij} + b'(\theta_{ij})) \frac{\partial \eta_{ij}}{\partial \theta_{ij}}. \end{aligned}$$

For  $g(x) = \log x$ , we have  $\frac{\partial \eta_{ij}}{\partial \theta_{ij}} = \frac{\phi_j}{\phi_j + \exp \eta_{ij}}$  and implies  $\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \eta_{ij}} = \phi_j \left( \frac{-y_{ij} + \exp \eta_{ij}}{\phi_j + \exp \eta_{ij}} \right)$ . Jointly we represent

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\eta}} = -\mathbf{Y} \circ \mathbf{B}^1(\boldsymbol{\eta}) + \mathbf{B}^2(\boldsymbol{\eta}),$$

where  $\mathbf{B}^1(\boldsymbol{\eta}) = [b_1(\eta_{ij})]_{n \times q}$  with  $b_1(\eta_{ij}) = \frac{\phi_j}{\phi_j + \exp \eta_{ij}}$  and  $\mathbf{B}^2(\boldsymbol{\eta}) = [b_2(\eta_{ij})]_{n \times q}$  with  $b_2(\eta_{ij}) = \frac{\phi_j \exp \eta_{ij}}{\phi_j + \exp \eta_{ij}}$ . Then, we compute  $\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\beta}} = \mathbf{Z}^T \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\eta}}$  and  $\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \mathbf{C}} = \mathbf{X}^T \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\eta}}$ . When  $\mathbf{C} = \mathbf{d}\mathbf{u}\mathbf{v}^T$ , we have  $\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \mathbf{u}} = \mathbf{X}^T \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\eta}} \mathbf{v}$  and  $\frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \mathbf{v}} = \left( \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\eta}} \right)^T \mathbf{X}\mathbf{u}$ .

## 1.7 Simulation

We report the model performance in case of  $p = 100$  in Table S1.

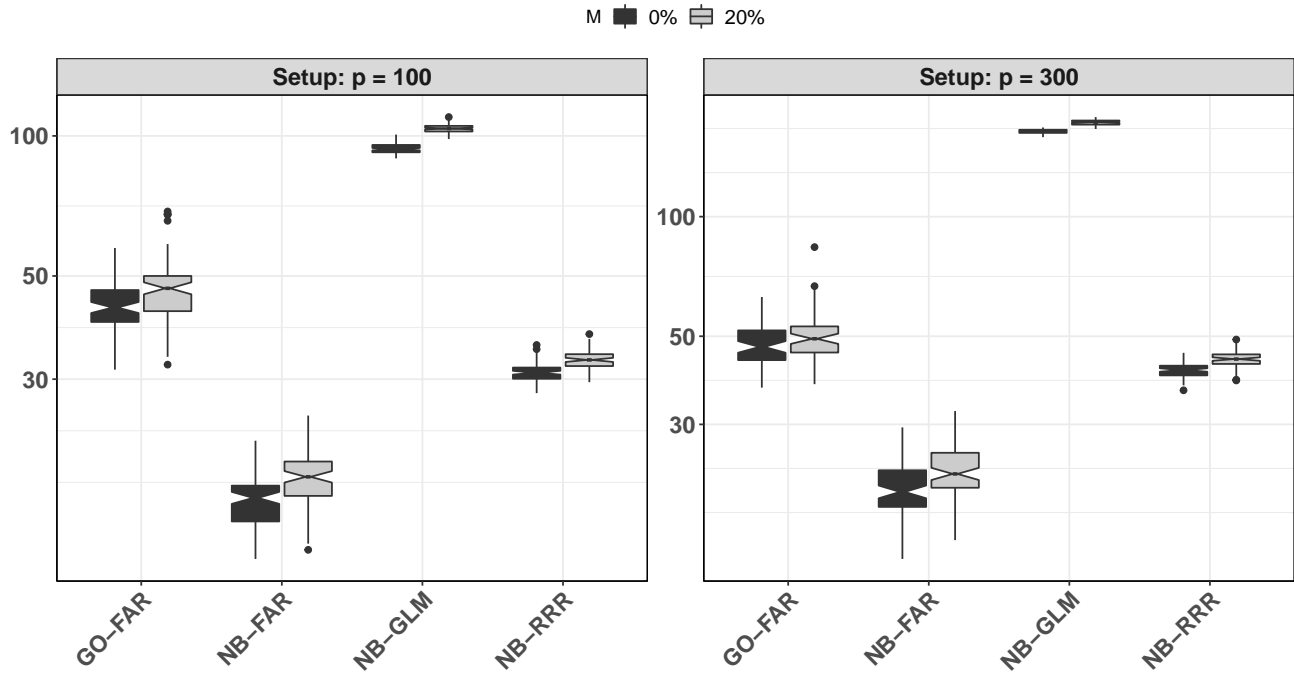


Figure S1: Simulation: notched boxplots of the estimation error  $Er(\eta)$  based on 100 replications in cases of  $p = 100$  and  $p = 300$ .

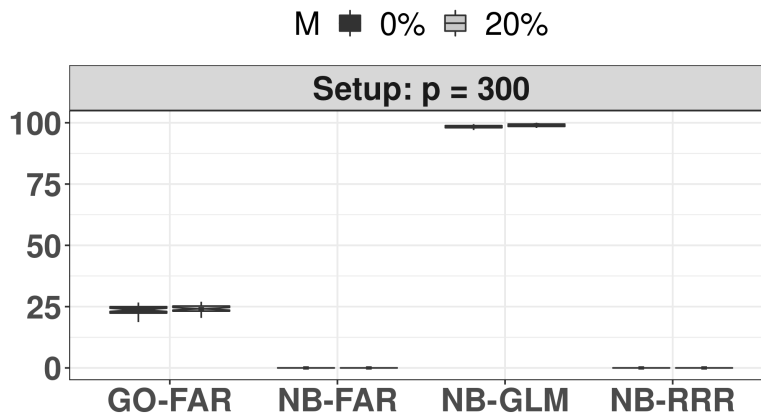


Figure S2: Simulation: notched boxplots of false positive rate, i.e., type I error, (y-axis) in a setting where true coefficient  $C^* = \mathbf{0}$  (null model). The analysis confirms NB-FAR's superior performance.

## 1.8 Application

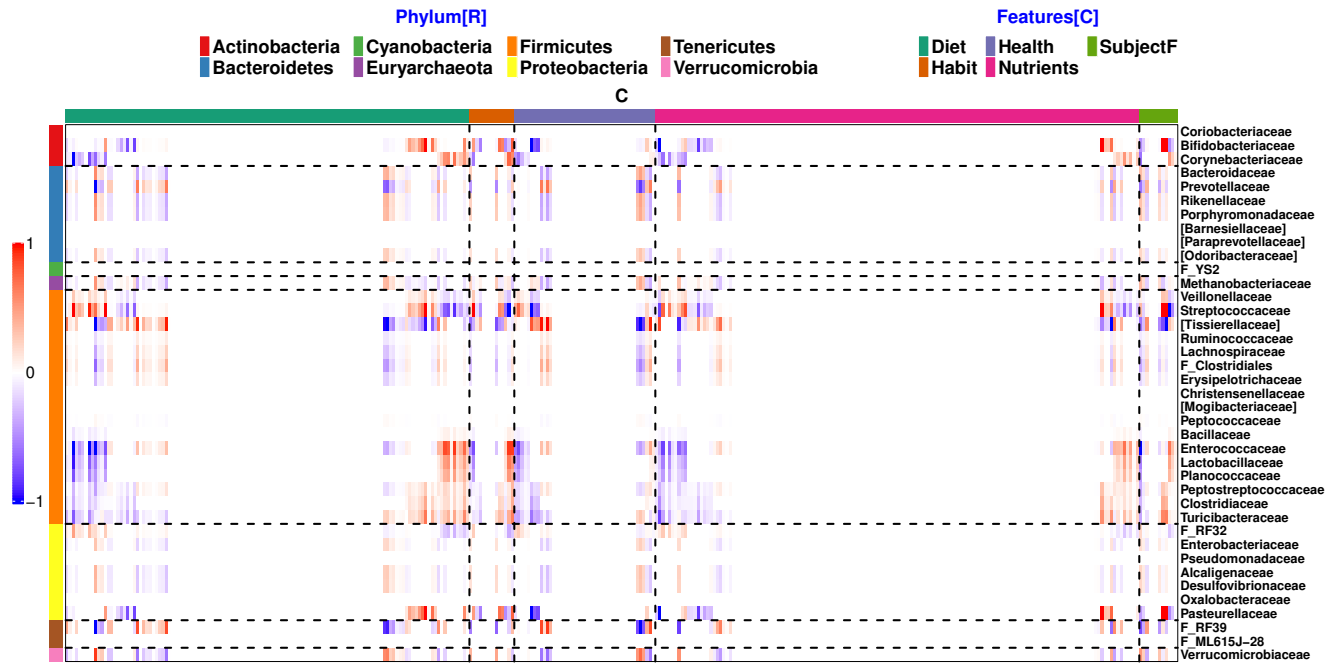


Figure S3: Application – AGP: The sparse estimate of the coefficient matrix  $\hat{C}$  using NB-FAR. Based on the Phylum of the taxon, horizontal lines separate the response into 7 categories given by Actinobacteria, Bacteroidetes, Euryarchaeota, Firmicutes, Proteobacteria, Tenericutes and Verrucomicrobia (top to bottom). Based on the type of the covariates, vertical lines (left to right) separate the selected predictors into five categories: namely, diet, habit, health, nutrients and subject features.

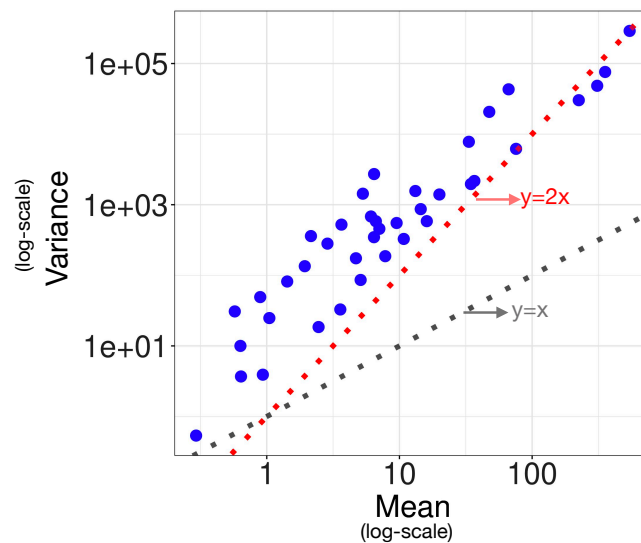


Figure S4: Application – AGP: Comparing mean and variance of the microbial abundance of each of 39 OTUs at the family level in the taxonomy to showcase overdispersed nature of the data.

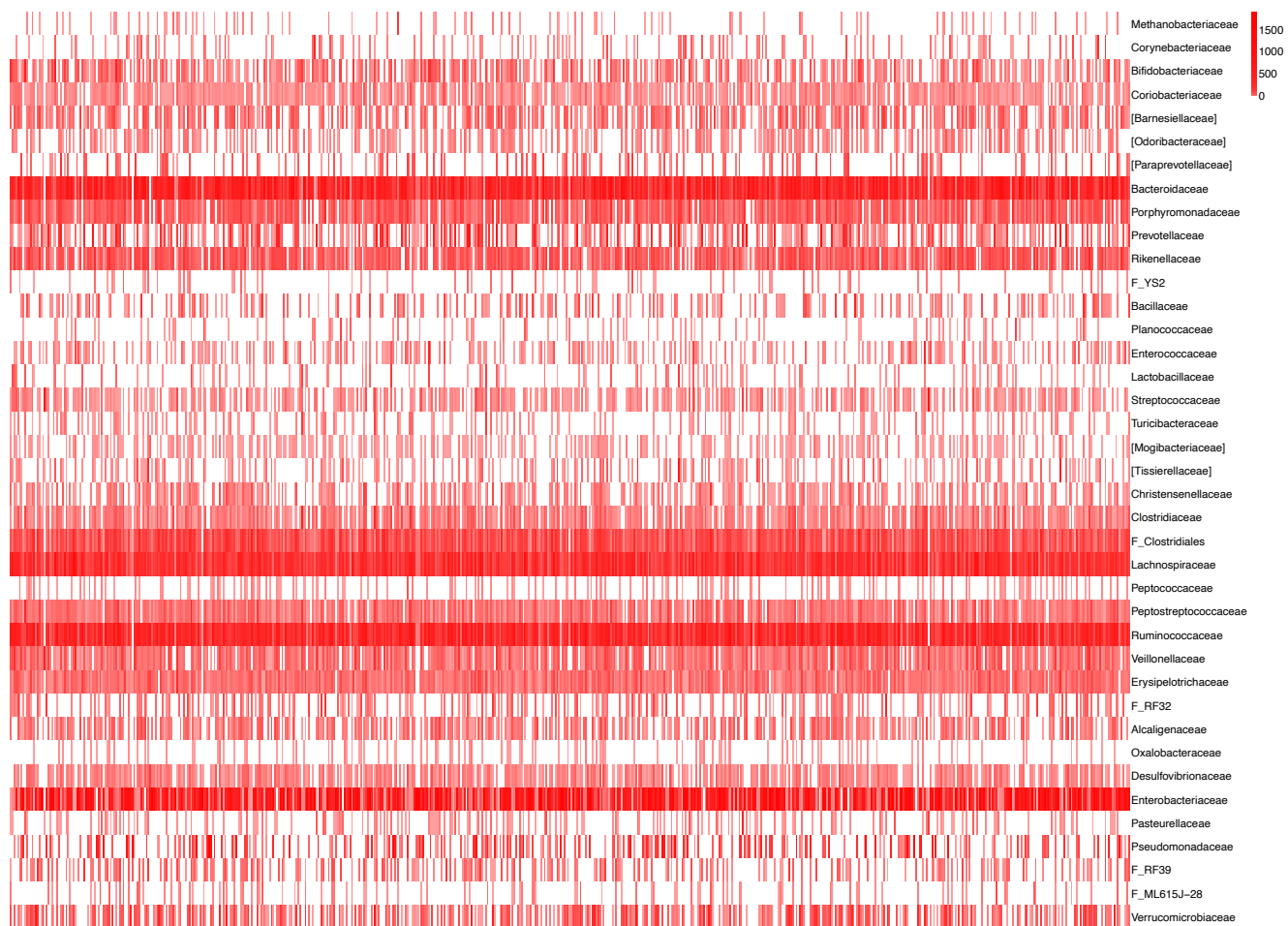


Figure S5: Application – AGP: Heatmap showing the abundance of each of the 39 OTUs at the family level in the taxonomy.



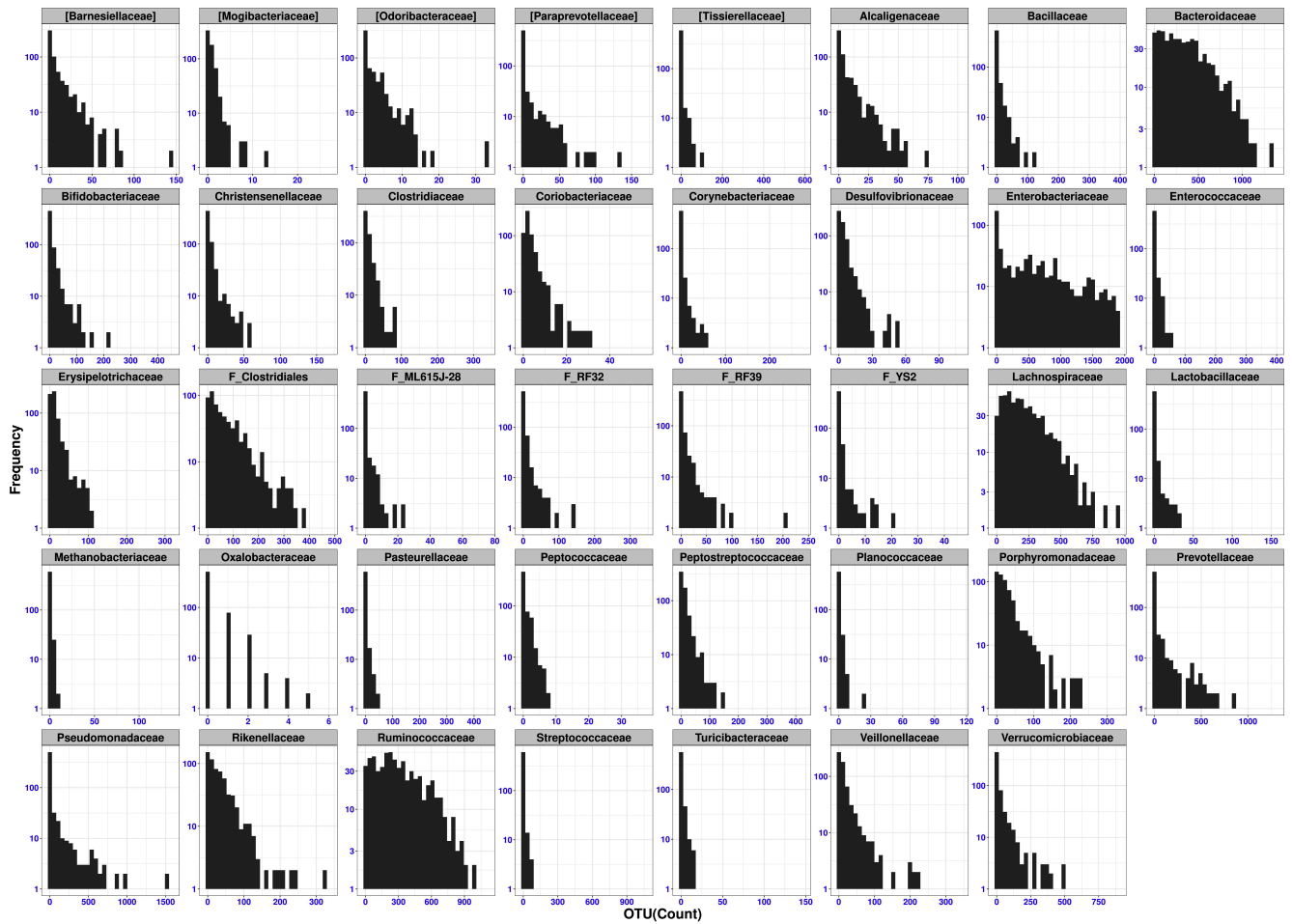


Figure S6: Application – AGP: Histogram showing the distribution of each of the 39 OTUs at the family level in n = 627 samples.

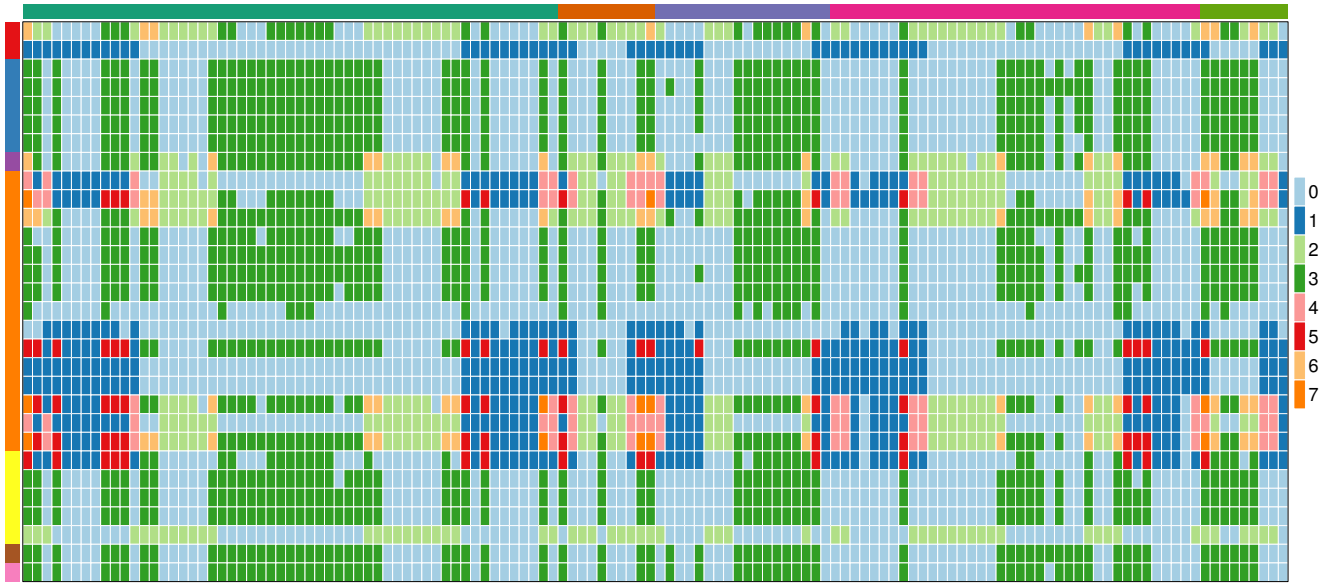


Figure S7: Application – AGP: Color code the contribution of the three unit-rank components in the sparse estimate of the selected rows and columns of the coefficient matrix  $\hat{\mathbf{C}}$ . Color code: a) 0 for no contribution, b) 1 for component  $k = 1$ , c) 2 for component  $k = 2$ , d) 3 for component  $k = 3$ , e) 4 for component  $k = \{1, 2\}$ , f) 5 for component  $k = \{1, 3\}$ , g) 6 for component  $k = \{2, 3\}$ , h) 7 for component  $k = \{1, 2, 3\}$ .

## References

- Noriszura Ismail and Abdul Aziz Jemain. Handling overdispersion with negative binomial and generalized poisson regression models. In *Casualty actuarial society forum*, volume 2007, pages 103–58. Citeseer, 2007.
- Chongliang Luo, Jian Liang, Gen Li, Fei Wang, Changshui Zhang, Dipak K Dey, and Kun Chen. Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis*, 167:378–394, 2018.
- Aditya Mishra, Dipak K Dey, Yong Chen, and Kun Chen. Generalized co-sparse factor regression. *Computational Statistics & Data Analysis*, page 107127, 2020.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL <https://www.R-project.org/>.
- Hui Zou and Trevor J. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x.